NeuroImage 178 (2018) 183–197

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

# A computational framework for the detection of subcortical brain dysmaturation in neonatal MRI using 3D Convolutional Neural Networks

Rafael Ceschin [a,b,*], Alexandria Zahner [b], William Reynolds [b], Jenna Gaesser [c], Giulio Zuccoli [b], Cecilia W. Lo [d], Vanathi Gopalakrishnan [a], Ashok Panigrahy [a,b]

[a] Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[b] Department of Radiology, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA, USA
[c] Division of Neurology, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA, USA
[d] Department of Developmental Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

ARTICLE INFO

*Keywords:*
Deep learning
Congenital heart disease
Neonatal imaging
Structural MR

ABSTRACT

Deep neural networks are increasingly being used in both supervised learning for classification tasks and unsupervised learning to derive complex patterns from the input data. However, the successful implementation of deep neural networks using neuroimaging datasets requires adequate sample size for training and well-defined signal intensity based structural differentiation. There is a lack of effective automated diagnostic tools for the reliable detection of brain dysmaturation in the neonatal period, related to small sample size and complex undifferentiated brain structures, despite both translational research and clinical importance. Volumetric information alone is insufficient for diagnosis. In this study, we developed a computational framework for the automated classification of brain dysmaturation from neonatal MRI, by combining a specific deep neural network implementation with neonatal structural brain segmentation as a method for both clinical pattern recognition and data-driven inference into the underlying structural morphology. We implemented three-dimensional convolution neural networks (3D-CNNs) to specifically classify dysplastic cerebelli, a subset of surface-based subcortical brain dysmaturation, in term infants born with congenital heart disease. We obtained a 0.985 ± 0. 0241-classification accuracy of subtle cerebellar dysplasia in CHD using 10-fold cross-validation. Furthermore, the hidden layer activations and class activation maps depicted regional vulnerability of the superior surface of the cerebellum, (composed of mostly the posterior lobe and the midline vermis), in regards to differentiating the dysplastic process from normal tissue. The posterior lobe and the midline vermis provide regional differentiation that is relevant to not only to the clinical diagnosis of cerebellar dysplasia, but also genetic mechanisms and neurodevelopmental outcome correlates. These findings not only contribute to the detection and classification of a subset of neonatal brain dysmaturation, but also provide insight to the pathogenesis of cerebellar dysplasia in CHD. In addition, this is one of the first examples of the application of deep learning to a neuroimaging dataset, in which the hidden layer activation revealed diagnostically and biologically relevant features about the clinical pathogenesis. The code developed for this project is open source, published under the BSD License, and designed to be generalizable to applications both within and beyond neonatal brain imaging.

## Introduction

Deep neural networks, or deep learning, are a set of machine learning algorithms that use nested layers of linear combinations of the original input allowing for the approximation of highly complex non-linear functions (LeCun et al., 1998). This property can be used in both supervised learning for classification tasks and unsupervised learning to derive complex patterns from the input data.

Deep neural networks have recently gained traction across a variety of domains, but none more than in imaging and computer vision. Naturally, the application of deep neural networks to clinical inference, biomarker discovery, and automated diagnosis in neuroimaging presents innumerable opportunities. Variants of deep learning have already been used extensively in neuroimaging applications, primarily in intensity-based classification tasks. Kleesiek et al. successfully implemented a 3-D Convolutional Neural Network (CNN) that outperforms the state-of-

[*] Corresponding author. Department of Radiology, Children's Hospital of Pittsburgh of UPMC, 4401 Penn Ave., Pittsburgh, PA, 15224, USA.
*E-mail address:* Rafael.Ceschin@pitt.edu (R. Ceschin).

https://doi.org/10.1016/j.neuroimage.2018.05.049
Received 3 April 2018; Received in revised form 4 May 2018; Accepted 19 May 2018
Available online 21 May 2018

the-art algorithms in skull stripping, generalizing well to multi-modal inputs including contrast-enhanced images (Kleesiek et al., 2016). Brosch et al. used 3-D CNNs to segment white matter lesions in brain MRI from patients with multiple sclerosis (Brosch et al., 2016). Gupta et al. used sparse auto-encoders and CNNs to detect lesions in structural MRIs using a large public dataset of patients with Alzheimer's disease (ADNI) (Mueller et al., 2005; Gupta et al., 2017). Payan et al. using a similar technique, but with 3-dimensional CNNs, was able to marginally improve the classification accuracy (Payan and Montana, 2015). Housseini-Asl et al. implemented 3-D convolutional auto-encoders to achieve excellent classification results using multi-modal imaging in this same dataset (Hosseini-Asl et al., 2016). Rajpurkar et al. (2017) achieved comparable results to expert radiologists on diagnostic chest x-ray images, and were able to localize the regions of the image that most contributed to the disease classification. For a comprehensive review of deep learning applications in a variety of multi-modal imaging on datasets including patients with Alzheimer's disease, schizophrenia, and various mild cognitive impairments, please see the review by Vieira et al. (2017) Of note, there is little data supporting the use of deep learning techniques in the neonatal neuroimaging domain, mainly due to small testable sample sizes and relatively increased complexity of brain structure.

As deep learning matures in this domain, CNNs have become an increasingly dominant approach to achieving great classification accuracies in the setting of small samples size, given appropriate constraints (Wagner et al., 2013), and relatively increased complexity of brain structures as seen in neonatal neuroimaging (Brosch et al., 2016). CNNs are a specialized variant of neural networks that leverages the inherent structural information encoded in images. Instead of indiscriminately using the entire image as the input to a given neuron, CNNs look for localized patterns across the image, recording the spatial location of an individual feature encoded by each neuron. This has two major advantages over traditional neural networks. First, CNNs decrease computational complexity, resulting in less encoding of overall parameters. Second, CNNs incorporate the spatial location of shared features within the input data. Taken together, these features open the possibility for the application of this technique to more limited, sparse datasets as seen in the setting of neonatal neuroimaging. In addition, 3D-CNNs have characteristic properties that make them exceptionally suitable for the morphological characterization of subtle structural differences, as seen in neonatal neuroimaging, particularly brain dysmaturation.

Neonatal brain dysmaturation refers to aberrant brain development in infants at risk for perinatal brain injury including those born prematurely or with congenital heart disease (Volpe, 2014). Neonatal brain dysmaturation can be classified into two broad types based on neuroimaging structural acquisition: reduced volume and/or shape disturbances (Back and Miller, 2014). Currently, most patterns of neonatal brain dysmaturation have been characterized by reduced volume, which can be the result of hypoplasia (lag in development) and/or a destructive mechanism multi-factorial in etiology (i.e. genetic, environmental or acquired injury). In contrast, subtle shape abnormalities can be relatively more difficult to visually assess and quantify, requiring either a trained expert in neonatal neuroimaging and/or sophisticated post-processing methods that are difficult to implement clinically. Volumetric information alone is not sufficient for accurate characterization. We have recently described a subset of term neonates with congenital heart disease that demonstrates brain dysplasia in subcortical structures including the olfactory bulb, cerebellum and hippocampus, as an example of brain dsymaturation related to visual shape disturbances (Panigrahy et al., 2016). The application of an automated machine learning technique, like CNNs, to detecting these types of neonatal brain dysplasia while also differentiating from volumetric abnormalities, would not only help facilitate basic discovery research related to etiological underpinnings, but also improve clinical detection for individualized patient care.

Here, we pair CNNs with previously developed neonatal structural brain segmentation methods to overcome some of the technical and biological limitations described above. We attempt to overcome limitations related to small sample size by first extracting each brain substructure of interest individually and transforming them into a standard space. This greatly reduces the search space required to learn the subtle abnormalities associated with a given pathology, making it feasible to implement a 3-D CNN as the classification algorithm. Additionally, enforcing spatial localization in the input data by pre-registering the structures into a standard space results in the feature maps retaining their spatial relation to the original structural morphology. The benefit of this approach is two-fold. First, once the model is derived from proper tuning of its parameters based on the training data, its implementation as a classification tool is straightforward, providing a feasible application of automated diagnostic systems in medical imaging. Second, by registering the input substructures into a standard space, the algorithm generates human-interpretable class activation maps of the hidden layers learned by the network, thereby retaining the original 3-D structural relationships. This refined algorithm gives us a data-driven model of the features within the dataset that contribute to the final classification, providing further insight into the structural morphology associated with the classification criteria and underlying pathology. We chose the cerebellum as the test substructure to test our algorithm because: (1) its relatively simple structure to segment compared to other regions of the brain, including the cerebral cortex and other subcortical structures including the hippocampus; (2) there are known neuroradiological criteria for cerebellar dysplasia in the developing human infant (Poretti and Boltshauser, 2016; Doherty et al., 2013; Oegema et al., 2015); and (3) anatomic related-sub regions of the developing cerebellum are important mediators of both genetics factors (Wegiel et al., 2010; Haldipur et al., 2017), and downstream pathways associated with neurodevelopmental impairment (Bosemani and Poretti, 2016; Stoodley and Limperopoulos, 2016). We tested our algorithm on a dataset of term neonates born with CHD at high risk for brain dysmaturation, including recently described cerebellar abnormalities, including both hypoplasia and dysplasia (Panigrahy et al., 2016).

## Materials and methods

### Subjects

We prospectively recruited 90 term-born neonates with congenital heart disease and 40 term-born healthy controls at Children's Hospital of Pittsburgh of UPMC with parent consent, as part of an ongoing Institutional Review Board approved study. Infants were scanned at close to term equivalent age, or when deemed clinically stable. 54 infants in the CHD cohort were scanned prior to their first surgical intervention, with the remaining 36 scanned post-surgical intervention. Infants were scanned on a 3T Siemens Skyra MRI (Siemens, Erlangen, Germany) without sedation using a 32-channel head coil. 18 infants in the control group were scanned on a 3T GE HDXT. Only infants who completed the volumetric imaging portion of the protocol were included in this study, and a cut-off of 52 weeks post-menstrual age was used to control for age-related morphological changes. Imaging parameters were as follows: (1) volumetric T1 Magnetization-Prepared Rapid Gradient-Echo at echo time (TE)/repetition time (TR): 418/3100 ms, $1.0 \times 1.0 \times 1.0\,\text{mm}^3$, and matrix size $320 \times 320$; (2) volumetric T2 Sampling Perfection with Application optimized Contrasts using different flip angle Evolution sequence at TE/TR: 2.56/2400 ms, $1.0 \times 1.0 \times 1.0\,\text{mm}^3$, and matrix size $256 \times 196$; (3) axial T2 Weighted Fast Spin Echo (FSE) at TE/TR: 3/2400 ms, slice thickness 2.0 mm with 0 skip, and in-plane matrix resolution $200 \times 256$.

### Dysplasia classification

We have previously described a pattern of dysmaturation in a subset of this population (Panigrahy et al., 2016). Each neonate's MRI was reviewed by an expert neuroradiologist blinded to their clinical history and classified as “normal” or “dysplastic” following existing qualitative imaging criteria developed by Barkovich et al. and others (Poretti and Boltshauser, 2016; Doherty et al., 2013; Oegema et al., 2015; Barkovich and Raybaud, 2012). The cerebellar dysplasia abnormalities are characterized by diffuse shape disturbance relative to brainstem/supratentorial structures and by abnormal orientation of the cerebellar fissures. Although the cerebellar vermis and cerebellar lobes were scored independently, we binarized the presence of dysplasia if it involved either cerebellar hemisphere or vermis. Infants identified to have hypoplastic substructures, but no evidence of structural dysplasia were not classified as abnormal in this study. For a subset of patients, we had two senior neuroradiologist score the brain dysplasia as previously described with a kappa score ranging between .86 and .91 for subcortical structures (Panigrahy et al., 2016).

### Experimental design

The full framework is summarized in Fig. 1. The pipeline takes as input the neonatal volumetric MRI, and segments 50 individual brain substructures. These substructures can be independently used for volumetric analysis. The substructure chosen for the classification task is then linearly registered onto a standard space to remove any size confounder, while retaining the shape information. This is the input to the neural network (Fig. 2). The initial architecture of the model to be evaluated is heuristically determined, followed by training and measurement of its performance on a small, randomly selected subset (30%) of the full dataset to prevent overfitting. This is iteratively performed, with incremental changes to the hyperparameters of the model, until satisfactory results are achieved for training on the full dataset with a fixed set of hyperparameters. This step is necessarily separate from validation of the architecture's final performance to prevent overfitting (Varoquaux et al., 2017). The final performance evaluation of the architecture is measured by cross-validation. While we did not perform nested cross-validation due to the limited data size and disproportionate amount of normal and dysplastic subjects, we minimized hyper-parameter tuning by starting with a heuristically determined architecture and restricting the epoch size of each parameter testing iteration. No hyperparameters were modified during cross-validation, and the accuracies reported are using the previously fixed architecture, measured on newly sampled partitions of the dataset. The output of the framework is a classifier able to detect structural dysplasia, and the hidden layers of the chosen network can be used to infer the morphological properties that contribute to the final classifier.

### Substructure segmentation

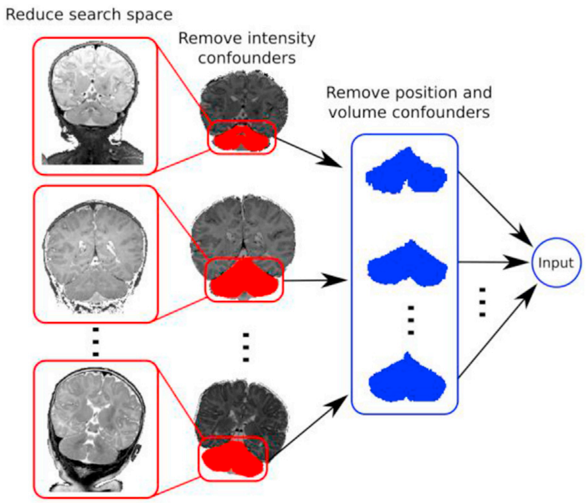
Reduce search space
Remove intensity confounders
Remove position and volume confounders
Input

**Fig. 2.** Dimensionality reduction of neonatal structural MRI decreases the search space by removing intensity, positional, and size confounders.

Each infant's cerebellum is extracted using a semi-automated processing pipeline developed in-house. This pipeline is written in Python using the Nipype framework (Gorgolewski et al., 2011), and the code is freely available at www.github.com/PIRCImagingTools/NeBSS. Fig. 3 shows an overview of the pipeline. The preferred input is the neonate's volumetric T2 image. However, this acquisition tends to be susceptible to motion artifacts, rendering much of the data inadequate for segmentation. As an alternative, we can register the T2 FSE images to the volumetric T1 3D images, utilizing the better T2 tissue contrast while still retaining accurate volumetric information. We first run a brain extraction with FSL's Brain Extraction Tool (BET) (Jenkinson et al., 2012), followed by FSL's bias correction algorithm (Zhang et al., 2001; Vovk et al., 2007) to remove any intensity gradient artifact due to field inhomogeneity. We then use the ALBERT neonatal parcellation dataset (Gousias et al., 2012, 2013) as the source template to propagate onto the subject space 50 discrete brain substructures modeled by the atlas using the Advanced Normalization Tools (ANTS) algorithm (Avants et al., 2009, 2011). ANTS calculates a symmetric, non-linear transformation between the target image and the source image. We choose as source four ALBERT subjects closest in gestation age to our subject to increase our registration accuracy and yield segmentations that more closely match the developmental stage of the subject. We then perform a voxel-wise winner-takes-all calculation across the four transformed ALBERT subject labels to determine the structure classification at each voxel of our subject. If there is a tie, the classification is randomly selected. We are left with 50

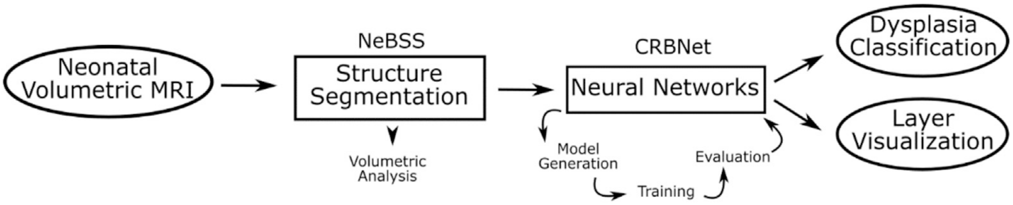

**Fig. 1.** Experimental Design and overview of the framework. The pipeline takes as input the neonatal volumetric MRI, and segments 50 individual brain substructures. These substructures can be independently used for volumetric analysis. These structures are then registered onto a standard space to remove any size confounder but retain the shape information. This is the input to the neural network. The architecture of a given network is heuristically determined, followed but training and evaluation of its performance. This is iteratively performed until satisfactory results are achieved. The output of the framework is a classifier able to detect structural dysplasia, and the hidden layers of the chosen network can be used to infer the morphological properties that contribute to the final classifier.

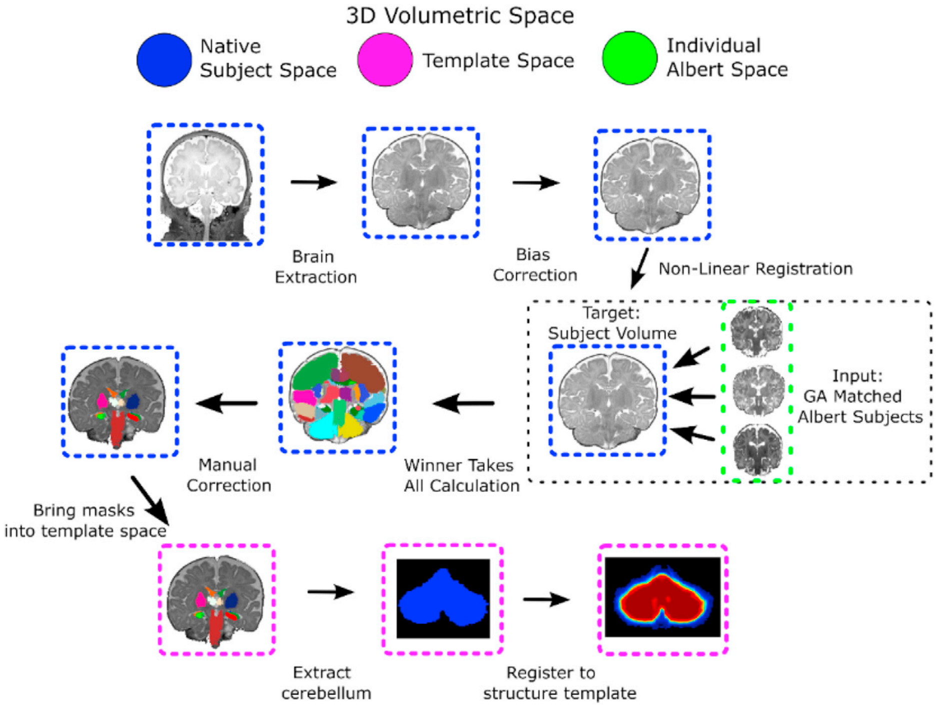

**Fig. 3.** Neonatal cerebellum structure extraction pipeline. We use an atlas based image registration pipeline to delineate the desired brain structures, which become the input to the classification task.

non-overlapping discrete regions mapped onto our subject space. This is followed by manual correction step to ensure anatomical accuracy of our label propagation. Finally, the cerebellum is linearly registered into a common template space, effectively removing any confounding volumetric information, but retaining the shape features of the structure. This serves the purpose of narrowing the search space and better enhancing the subtle morphological variations across the subjects.

As the output of our pipeline is highly dependent on the accuracy of the substructure extraction, we tested our reproducibility using the Dice Similarity Coefficient (DSC) (Zou et al., 2004) between two images:

$$DSC = \frac{2|A \cap B|}{|A| + |B|}$$

We assessed the inter-rater reliability by having two independent users perform manual correction on the same set of six subjects (3 Controls and 3 CHD). For intra-rater reliability, one user performed a repeated round of corrections on the same set of subjects.

### Volumetric analysis

To first investigate whether the volumetric information extracted from our pipeline is sufficient for accurate prediction of structural dysplasia, we tested three logistic regression models, and evaluated their accuracy using Leave-One-Out cross-validation (LOOCV). LOOCV iterates through the dataset, removing one sample from the set, and using the remaining samples to train a classifier. The classifier is then tested on the left out single sample. The averaged classification accuracy for each model is then calculated. This method of validation is designed to reduce overfitting of a classifier primarily by reducing its dependence on outliers, while still retaining a large training set for learning. The first model we tested was as follows:

$$\log it(Cerebellar\ Dysplasia) = \beta_0 + \beta_1 PMA + \beta_2 Cerebellar\ Volume + \beta_3 CHD + e$$

Where PMA is the post-menstrual age at time of scan, cerebellar volume is the manually corrected volume extracted from the pipeline, and CHD is the subject's CHD/control classification. This model included the entire dataset. Additionally, we tested this model without the CHD classification term to better represent a naïve classifier agnostic to the patient's clinical diagnosis, as well as training only on the subjects with CHD to minimize noise from healthy controls, as only patients in the CHD cohort were identified to have cerebellar dysplasia.

### Neural networks

A neural network is comprised of units (neurons) that apply an arbitrary, non-linear activation function **σ** to a linear combination of the inputs and a set of learned weights and bias:

$$a = \sigma(w^T x + b)$$

where w is a vector of weights, x is the input and b is the bias term. In a deep neural network, the output of each layer is fed into the next layer. Any given layer in a deep neural network can have an arbitrary number of neurons, each learning their own weights and bias, allowing for a high-dimensional approximation of any non-linear function, given enough neurons and layers. Learning the weights and biases of a neural network is done by an optimization algorithm known as Stochastic Gradient Descent (SGD). The first requirement in gradient descent is the selection of a cost function. This cost function is simply a measure of the accuracy of our classification task. Common cost functions used in neural networks include mean squared error (MSE), negative log-likelihood, and cross-entropy. The negative log-likelihood cost function is particularly powerful in conjunction with a final softmax activation layer:

$$\sigma(z)_i = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}\ for\ j = 1, \ldots, K$$

where z is the weighted sum of the outputs of the previous layer, and K is the total number of neurons in the final layer (the classifiers). Softmax normalizes the output of each neuron of the final layer in the network over all possible output neurons, which are the desired classifiers. The output of the softmax for each neuron (j) in the final layer is then the likelihood of the given input being classified as that particular label. Therefore, minimizing negative log-likelihood function is equivalent to increasing the prediction accuracy of our network. Gradient descent works by calculating the gradient of the cost function given the current parameters of the network, and then changing these parameters in the opposite direction of the gradient, scaled by a small factor - called the learning rate.

SGD uses a simple update rule to calculate the new parameters V':

$$V \to V^{'} = V - \eta \nabla C$$

where V are the parameters of the network, $\eta$ is the learning rate, and $\nabla C$ is the gradient of the function given the current parameters. The gradient of the cost function is analogous to its slope, generalized to more than two dimensions, and indicates the direction in which the function is increasing. Thus, by making small incremental changes in the weights and bias opposite of their gradient, we iteratively minimize the cost function. Given a small enough learning rate and enough iteration, SGD is guaranteed to converge on a (local) minimum.

### Convolutional neural network

Traditional neural networks consist of fully connected layers, taking as input a flattened vector of the data. While this architecture has proven powerful across numerous domains, from genomics to text mining, they do not take advantage of the intrinsic spatial information contained in images. Such a network only learns patterns of activation as they appear in a fixed order of the training data. A proposed solution to this limitation is the use of Convolutional Neural Networks (CNNs). Instead of feeding the entire image into each neuron, a CNN convolves a filter of reduced size, the Local Receptive Field (LRF), with the input image. The result of this convolution is a set of neurons, which take as input only the regions of the input image within their respective LRF, but encode the spatial location of the feature in the original image space. This is able to identify position invariant, reusable features by having one full set of neurons in a hidden layer (called a feature map), share the same parameters. This has two advantages: (1) we greatly reduce the number of parameters needed to compute at each layer, and (2) the feature map is now a spatial representation of the features present in our data. Convolution layers are often followed by max-pooling layers, which take the maximum value of each parameter within the selected pooling filter. This effectively denoises the data while also reducing the dimensionality of the subsequent layer.

There is no closed form method of designing an optimal neural network. Collectively, the parameters that make up the architecture: learning rate, number of layers, feature maps per layer, regularization coefficient, LRF, and max-pooling size, are known as the hyper parameters of the model. The nature of deep neural networks makes it so that any given network is never optimal, as there exist an infinitely large number of network architectures that may perform equally or better. Instead, we strive to design an effective network, and iteratively try to improve on it. We can use heuristic approaches to expedite the search for ideal parameters and establish a reasonable starting point before committing to fully training a specific network architecture, which can take weeks or months depending on the size and complexity of the dataset. We can instead perform quicker cycles of learning on a subset of the data without reaching convergence or saturation. Here, we applied a combination of random and grid search (Zhao et al., 2015) to cycle over viable hyper parameters, giving us an estimation of what hyper parameters are likely to work.

Overtraining is always a concern with highly complex machine learning algorithms. Overtraining occurs when the model implicitly learns features specific to the training data, effectively “memorizing it”, but does not generalize to external datasets. To reduce these effects, we can impose restrictions on our cost function and hidden layers. We used two complementary approaches to overtraining prevention: L2 regularization and layer dropout. L2 regularization adds an additional term to the cost function:

$$\frac{\lambda}{2n}\sum_{w} w^2$$

where λ is an additional hyperparameter, n is the number of subjects in the training batch and w are the weights. This method adds a penalty for higher weight values, attenuating runaway effects that can lead to over fitting the data. Similarly, layer dropout is a method of ensuring more generalizability in the classification test. Layer dropout randomly removes a pre-determined number of neurons from the final fully connected layers at each training iteration. This prevents the learning algorithm from relying too heavily on any one neuron, enforcing more generalization distributed across the entire network space instead of localized neurons.

### Validation

To evaluate our model, we performed a 10-fold cross validation. This is done by partitioning the dataset into 10 independent sets, and at each iteration using 9 sets as the training set and the remaining data for validation of the classification accuracy. We trained each validation run for a total of 100 epochs. The best performing parameters from the validation runs were then used in a final learning run of 700 epochs to generate the activation maps described in the next section.

### Hidden layer visualization

Interpreting the features learned by a CNN could potentially provide insight into the biological and structural features that contribute to a given substructure's malformation. However, the parameters learned in the hidden layers of a neural network are traditionally treated as a black box. We have no direct control on the features learned at each layer, and it has been shown that each individual unit does not often hold any meaningful semantic information, but rather the combination of features within the entire space hold this higher level of abstraction (Szegedy et al., 2013). Several methods have been proposed in attempts to visualize the information contained in the hidden layers. Most notably, Zeiler and Fergus used a Deconvolutional Network (deconvnet) (Zeiler et al., 2010; Zeiler and Fergus, 2014) to project the hidden layer activations back into pixel space as a form of pre-training and quality control. This method generates a projection in pixel space that reflects the features at each layer that mostly contribute to the final classifier, independent of spatial location. This approach has shown to be beneficial, particularly when analyzing images with positionally varied features across samples.

In our work, we approximate this method by directly visualizing the mean activation maps at each hidden layer for each cohort. By implementing a 3-D CNN, we retain the volumetric structure of the input data, as each hidden layer's feature map will encode the local features within the input that contribute to the final classification. As each subject's cerebellum is linearly registered into a common space, we directly impose structural information into the input data. This allows us to project the mean layer activations for each group into a 3-D space and view them as a proxy for the anatomical features that lead to the final diagnosis. One drawback of this approach, however, is that this loses spatial information at each max pooling step, which can be particularly undesirable when using binarized, sparse data as input.

To further illustrate the benefits of spatially constrained filters, Fig. 4 shows synthetic examples of possible filters that can be learned by a CNN. Randomly distributed filters (Fig. 4-A) may still result in good classification accuracy, but provide no anatomically interpretable benefits. Random filters may also be symptomatic of an overfitting algorithm. Geometric filters (Fig. 4-B) are often seen in image classification algorithms with no structural coherence of the input data. These filters are typically found in earlier layers of deep networks and are thought to be

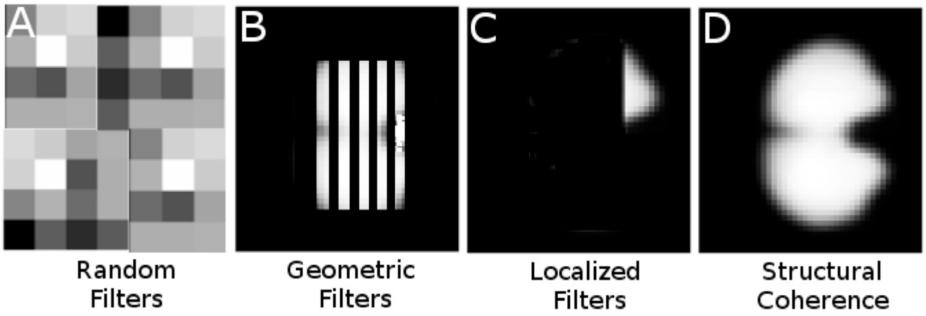

**Fig. 4.** Non-exhaustive, synthetic examples of possible learned filter outcomes when training CNNs. Randomly distributed filters (A) may still result in good classification accuracy, but provide no anatomically interpretable benefits. Geometric filters (B) are often seen in image classification algorithms with no structural coherence of the input data. These filters are typically found in earlier layers of deep networks, and are thought to be used in further layer abstractions to combine to form more complex features. Localized filters (C) can be useful for identifying specific regions of the input image that highly contribute to the final classification. Structurally coherent activation maps, as are generated by our methods due to structural constraints, (D) retain the structure of the input data, and selectively activate regions of the input dataset that contribute to the final classifier.

used in further layer abstractions to combine to form more complex features. While mechanistically interpretable, they also would not provide useful anatomical insight in this use case. Localized filters (Fig. 4-C), where only a subsection of the input image activates the filter, can be useful for identifying specific regions of the input image that highly contribute to the final classification. While potentially desirable due to their specificity, localized filters are not likely to be observed when classifying complex structural morphology but would theoretically be useful in lesion classification or other general discrete feature detection. Structurally coherent activation maps, as imposed as a constraint in this work (Fig. 4-D), retain the structure of the input data, and selectively activate regions of the input dataset that contribute to the final classifier. In this work, we attempt to leverage this property to identify complex patterns within the input structure that ultimately differentiate normal and dysplastic subtypes.

Building upon this concept, our final visualization approach aims to implement a similar method proposed by Zhou et al. (2015) by generating Class Activation Maps (CAMs) to spatially locate the regions of the input image that most contribute to the final classifier at the each convolutional layer. Zhou et al. used a global average pooling (GAP) step prior to the final classifier. This has the advantage of assigning one weight to a spatial map prior to the classifier, providing a measure of importance of the specific feature encoded by each feature map, at the cost of losing spatial coherence. However, by propagating this weight back onto the final 3D convolutional layer, they were able to create a weighted heat map of the spatial location of the most important features to the classifier for each training sample. For each training sample, the CAM is then defined as:

$$M_c(x,\ y) = \sum_k w_k^c f_k(x, y)$$

Where (x,y) are the downsampled pixel coordinates at the last convolutional layer of samples of class $c$, and $w_k^c$ is the final weight given to the kth feature map $f_k$ after the GAP step. This approach has proven powerful for spatially variant, multi-labeled classification problems. Here, since we have fully connected layers prior to the final classifier rather than a GAP layer directly connected to a classifier, we lose the direct connection of an individual feature map's contribution to the final classifier. However, we can still extract salient anatomical information by generating layer specific CAMs at each convolutional layer, for both normal and dysplastic structural variants, here denoted as within-class average activation maps (wCAMs).

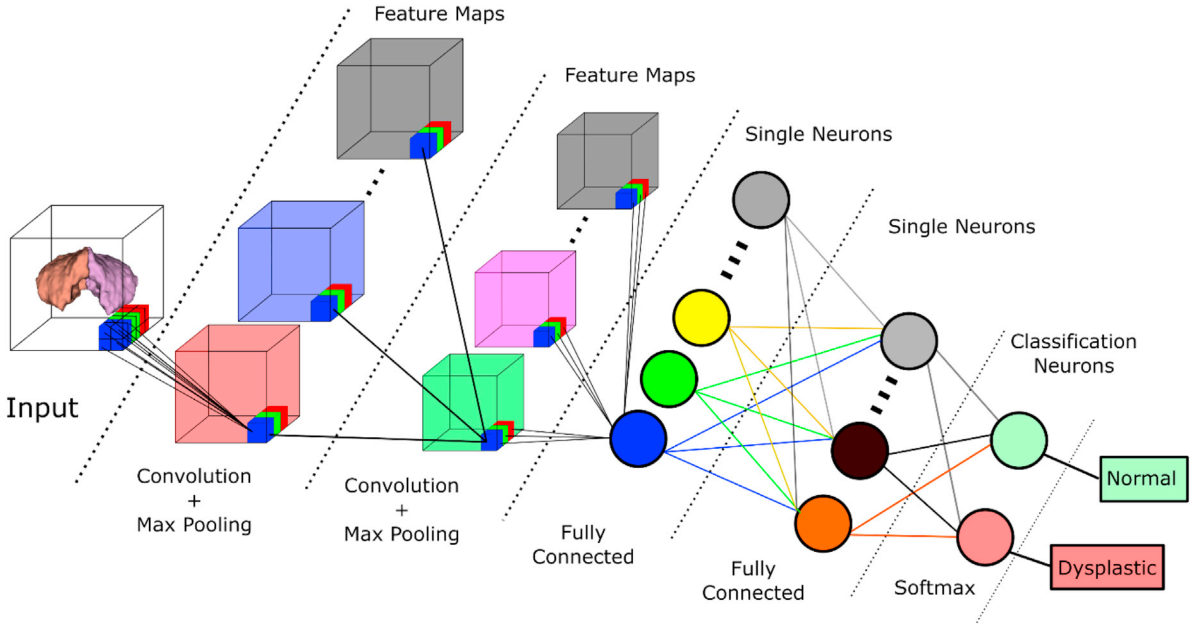

**Fig. 5.** 3-D Convolutional Neural Network Overview. Simplified architecture of the generated CNN. The algorithm takes as input the extracted, binarized cerebellum and outputs a classification of dysplastic or normal structure.

### Implementation

Fig. 5 shows a simplified overview of the general architecture used as the basis for the 3D CNN. The computational framework was developed using Python and the Theano (Theano Development Team, 2016; NielsenNeural Networ, 2015) library, with customized routines for the 3D convolution and neuroimaging processing. The pre-processing of the input to the network uses the Neuroimaging in Python (Nipy) library (Gorgolewski et al., 2011). The full code can be viewed at: www.github.com/PIRCImagingTools/CRBNet.

## Results

### Subjects

Mean gestational age was 38.0 weeks ( ±2.9) in the group of infants with CHD, and 41.2 weeks ( ±3.8) in the healthy control group. Mean post-menstrual age (PMA) at time of scan was 42.4 ( ±6.9) and 43.5 ( ±5.5) weeks for infants with CHD and healthy controls, respectively. Infants within the control group showed higher PMA-adjusted cerebellar volumes when compared to neonates with CHD, with a mean cerebellar volume 5201.5 ( ±7930.6) $mm^3$ larger than the dataset mean volume, with neonates with CHD showing on average 1035.4 ( ±5591.7) $mm^3$ smaller than the dataset mean volume ($p < 0.000$). We saw no statistically significant difference in PMA-adjusted cerebellar volume when comparing infants who had pre-op imaging and infants with post-op imaging, with pre-op MRI's having on average 941.51 ( ±3214.5) $mm^3$ larger cerebelli than the entire CHD cohort mean, and post-op MRI's, showing 1439.7 ( ±7720.9) $mm^3$ smaller cerebelli than the CHD cohort mean ($p < 0.053$). The cerebellar data included in this analysis showed no evidence of perinatal injury as we excluded cases with cerebellar infarcts and cerebellar hemorrhage.

The logistic regression approach was not able to discriminate structural dysplasia using age and volumetric information. While the patient's CHD classification yielded a very high odds ratio of $1.0 \times 10^8$ (as only patients with CHD had dysplastic cerebelli), it was not statistically significant ($p < 0.991$) and did not yield any discriminatory power. Cerebellar volume yielded odds ratios near 1.00 for all three models tested and was not statistically significant. Similarly, we saw no statistically significant difference in cerebellar volumes between subjects classified as structurally normal and subjects with dysplastic cerebelli ($p < 0.321$), shown in Fig. 6.

Among the infants with CHD, 17 (18.9%) were diagnosed with a dysplastic cerebellum. The total incidence rate in the entire dataset was 13.1%. To attenuate the effects of the low incidence of abnormal cases in the training set, we bootstrapped the data by randomly sampling from the set of abnormal cases and artificially inflating the dataset to contain a more proportional ratio of controls to abnormal cases, with a final training dataset containing 47% dysplastic structures. To prevent overtraining of the algorithm by oversampling the same small subset of cases, we introduce a small amount of random translation (3–5 voxels) in 3 directions to each case in the dataset. This helps prevent the introduction of fixed-position based artifacts and aids in generalization of the parameters, but still retains the spatial integrity of the input data.

**Table 1**
Intra- and Inter-rater reliability of manual corrections in structure extraction.

| | Structure | |
|---|---|---|
| | L Cerebellum | R Cerebellum |
| Measure | mean (sd) | mean (sd) |
| Inter-Rater | 0.943 (0.01) | 0.951 (0.02) |
| Intra-Rater | 0.952 (0.03) | 0.941 (0.03) |

### Substructure segmentation

Table 1 shows the Dice coefficients for inter- and intra-rater reliability of the manual corrections performed on a subset of infants. Users were blinded to the infant's cohort and dysplasia classification prior to correction. Measures were highly consistent across both inter- and intra-raters, with the lowest mean Dice coefficient of 0.941 ( ±0.03) between one rater measuring the right cerebellum and highest mean Dice coefficient of 0.951 ( ±0.02) between multiple raters measuring the same substructure.

### CNN parameters

Table 2 shows the final parameters for our chosen architecture. It consists of a total of 7 hidden layers, with 4 initial convolutional layers, followed by 2 fully connected layers and a final softmax classification layer. Each convolutional layer is followed by a max-pooling procedure. The initial learning rate (η) was set to 0.005 with a scheduled rate decay of 0.5*η every 40 epochs. We used the negative log-likelihood cost function, with an added L2 regularization hyperparameter (λ) set to 0.01. We used a layer dropout parameter of 0.3. Finally, we implemented a momentum update method with an initial μ value of 0.5, increased to 0.9 after a stabilization period of 15 epochs.

Modern computer vision CNNs have achieved excellent results and improved learning speed using the Rectified Linear Units function (ReLU), and more recently Exponential Linear Units (ELU) as the activation function (Clevert et al., 2015). However, we achieved rather mediocre results with ReLU in our application, with runs never achieving

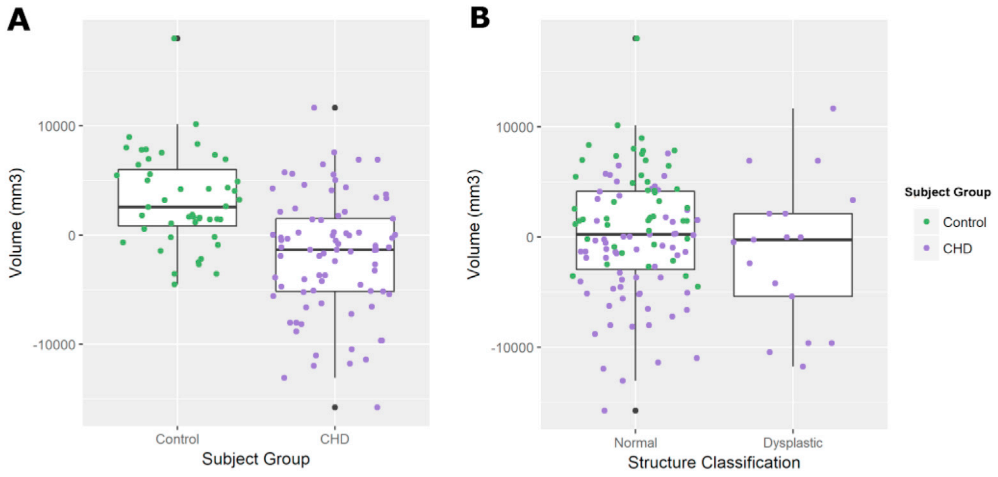

**Fig. 6.** Boxplots showing post-menstrual age corrected cerebellar for between A) control subjects and patients with CHD and B) Subjects classified as having structurally normal cerebelli and subjects diagnosed with dysplastic cerebelli. Only subjects with CHD were diagnosed with dysplastic cerebelli. There was a statistically significant difference in volumes between neonates with CHD and controls ($p < 0.000$) but no difference between dysplastic structures and normal structures ($p < 0.321$), suggesting that hypoplasia is independent from structural dysplasia.

**Table 2**
3-D CNN architecture.

| Layer | Layer Type | Input Dimension | # Neurons/ Feature Maps | LRF | Pooling |
|---|---|---|---|---|---|
| 1 | Convolutional/ Max Pool | 100 × 90 x 70 | 10 | 4 × 4 x 4 | 2 × 2 ×2 |
| 2 | Convolutional/ Max Pool | 48 × 43 x 33 × 10 | 15 | 2 × 2 x 2 | 2 × 2 ×2 |
| 3 | Convolutional/ Max Pool | 23 × 21 x 16 × 15 | 25 | 2 × 2 x 2 | 2 × 2 ×2 |
| 4 | Convolutional/ Max Pool | 11 × 10 x 7 × 25 | 50 | 2 × 2 x 2 | 2 × 2 ×2 |
| 5 | Fully Connected | 5 × 4 x 3 × 50 = 3000 | 300 | – | – |
| 6 | Fully Connected | 300 | 100 | – | – |
| 7 | Softmax | 100 | 2 | – | – |

convergence. This is likely a result of the sparse, binarized nature of the inputs, which can lead to exploding and/or vanishing gradients during training. Instead, our final architecture uses the tanh function:

$$\tanh = \frac{e^{2x} - 1}{e^{2x} + 1}$$

which has traditionally performed well in many classification tasks, at the cost of slower learning at the saturation extreme compared to newer activation functions such as ReLU and ELU.

### *Cross-validation*

Fig. 7 shows the mean cost at each epoch the 10 cross-validation runs. To ensure that the training and validation sets remain independent, the resampled subjects from the bootstrapping pre-processing remained within their own validation sets without crossover into the remaining dataset. While some variance is expected due to the random initialization of weights, all runs converge within 50 epochs. Fig. 8 shows the classification accuracy for each run. The training set is a bootstrapped dataset with an inflated incidence of abnormal cases. The data is partitioned into 10 independent sets, where at each run 9 sets are used to train the network and the final set is used as the validation set. The test set is the original dataset (without the bootstrapped cases added in). All runs achieved 100% classification accuracy in the training set within 50 epochs. The average classification accuracy on the validation set was 0.985 ± 0.0241, with several folds reaching 100% classification accuracy. To inspect the cross-validation results for any subject specific artifacts that may erroneously contributed to the interpretability of the hidden layers, we generated class activation maps for each convolutional layer of each cross-validation run. The wCAMs for layers 3 and 4 are shown in Supplemental Figs. 1 and 2. While some variance is observed in activation values outside of the cerebellar structure, particularly in layer 3, the contribution of the anterior and inferior regions of the cerebellum, as well as the vermis, are consistently present. In layer 4, some variance is observed in features constructed in the center of the structure across folds, but the features differentiating dysplastic and normal structures in the region of the vermis and cerebellar hemispheres is consistent across folds.

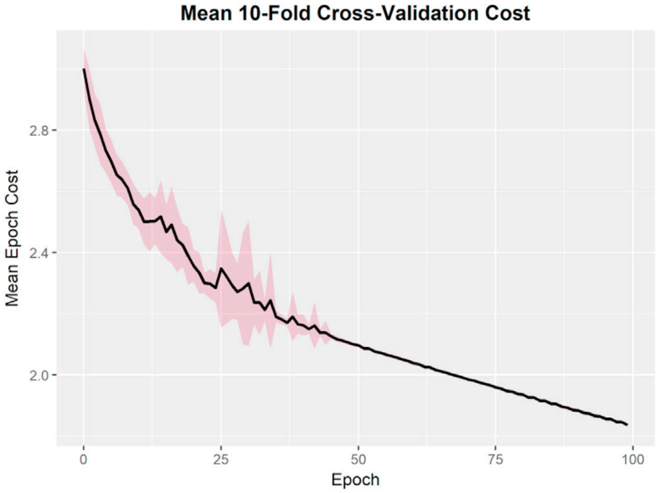

**Fig. 7.** 10-Fold Cross validation mean and standard deviation cost across all runs. Cost function was the negative log-likelihood function with an L2 regularization parameter of 0.01. While some variance is expected due to the random initialization of weights, all runs converge within 50 epochs.

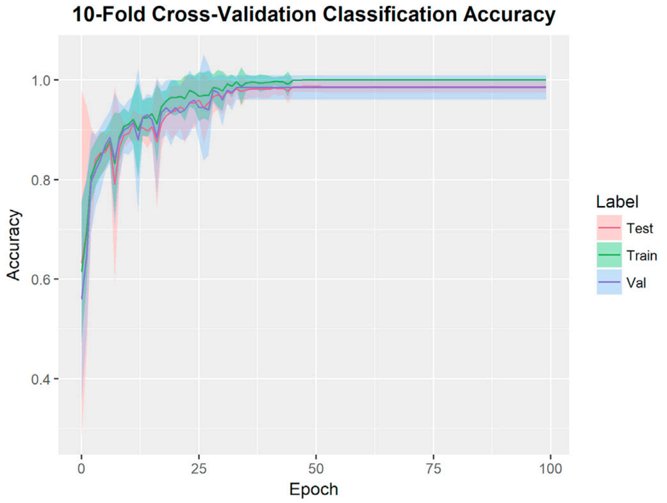

**Fig. 8.** 10-Fold Cross validation mean and standard deviation classification error for each dataset. The training set is a bootstrapped dataset with an inflated incidence of abnormal cases. The data is partitioned into 10 independent sets, where at each run 9 are used to train the network and the final set is used as the validation set. All runs achieved 100% classification accuracy in the test set within 50 epochs.

### *Visualization results*

The first layer's mean activations for the entire dataset is shown in Fig. 9. As the algorithm parameters are randomly initiated, there is no intrinsic information to the order of the learned filters; therefore the filters are selectively sorted for visual clarity. Intensities are scaled to show the contrast in activation range in each filter. We see that, as expected due to the structural coherence imposed by linear registration and binarized input, each filter distinctly delineates the cerebellum, with some filters showing higher activations limited to the perimeters of the structure. This corresponds to the cerebellar cortex of the bilateral cerebellar hemispheres, and serves as a potential edge detectors. Comparatively, the activations in layer 2 (Fig. 10) show more dramatic delineations of the superior surface of the cerebellum, co-localizing to the posterior subdivision of the bilateral cerebellar cortical hemispheres and also to the midline vermis. Note that filters with visually similar activations are not shown for clarity. Subsequent layers show a similar increase in the range of activations within each filter, however, as they are downsampled at each hidden layer, the features become less interpretable in an anatomical context.

We can then look at the difference in activations between the dysplastic and normal cohorts. The difference map of the first layer is shown in Fig. 11. Blue regions are areas in which we see higher activation in the dysplastic cerebelli, and red is increased in the normal substructures. White regions showed low activations equally in both cohorts, and are thus uninformative to the classification task. As with the previous

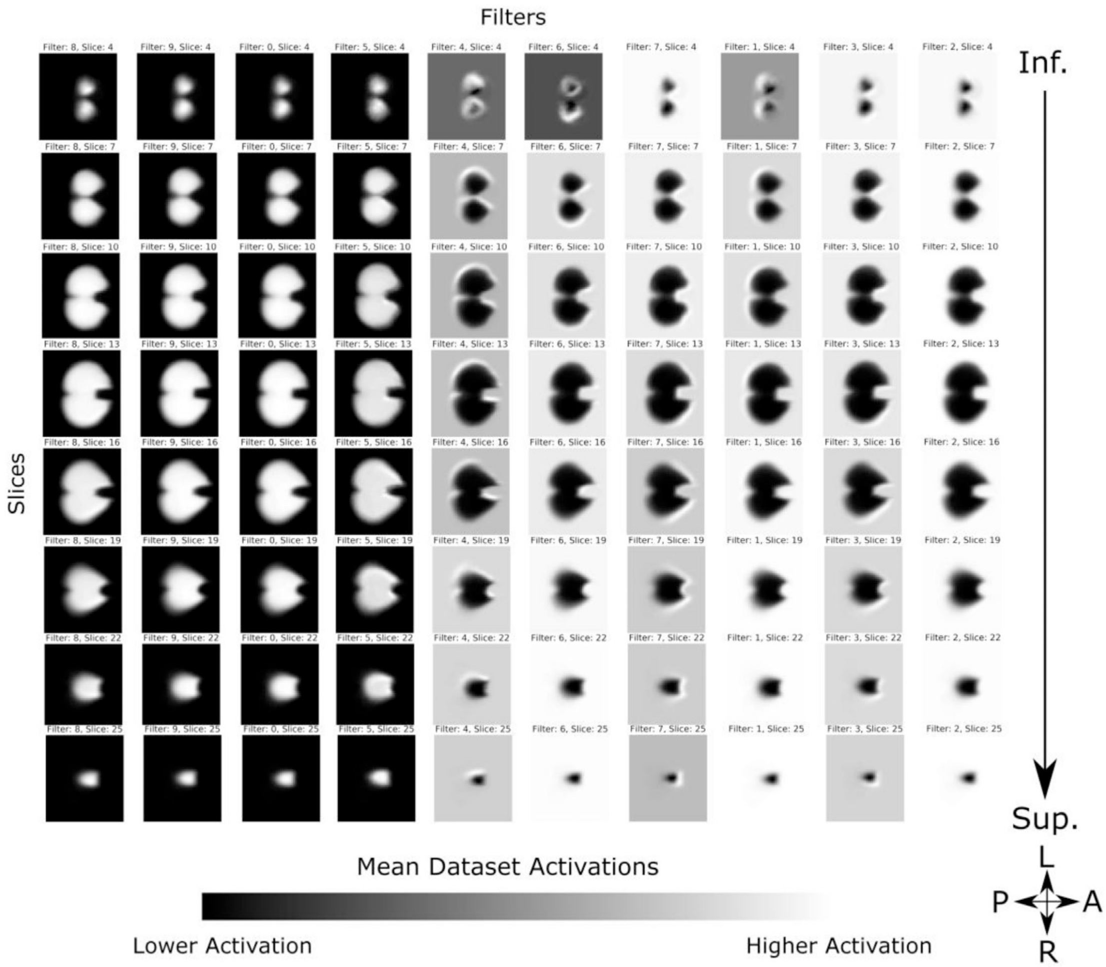

**Fig. 9.** First convolutional layer mean activations. Filters are selectively sorted for visual clarity and intensities are scaled to show the contrast in activation range in each filter. Each filter distinctly delineates the cerebellum, with some filters showing higher activations limited to the perimeters of the structure, co-localizing to the cerebellar cortex of the bilateral cerebellar hemispheres, serving as potential edge detectors.

figures, filters are sorted for visual clarity to separate the filters that show higher activations in normal cerebelli contrasted with dysplastic inputs. Each filter shows a predilection for primarily classifying either normal or dysplastic substructures, however, some overlap is observed. We see a clear pattern where the normal substructures show more defined delineation of the inferior and lateral surface of the cerebellar lobes while the dysplastic substructures show more defined superior surface of the cerebellum that anatomically correspond to the posterior lobe of the cerebellar hemispheres and the midline vermis. Similar to the mean activation maps, the first layer primarily serves the purpose of global delineation of the cerebellum. In the second layer's difference maps (Fig. 12), we see more regional delineation of key cerebellar regions that discriminate between dysplastic and normal cerebellar tissue. Specifically, the superior surface of the cerebellum (corresponding anatomically to the posterior lobe subdivision and the midline cerebellar vermis) shows significant discriminatory power, seen as strongly differentiated regions of the activation maps, between dysplastic and normal tissue structure. No lateralization difference (between right and left) was appreciated in the cerebellar hemispheres or midline vermis. In contrast, deep cerebellar regions (including the flocculonodular lobe subdivision) were not informative to the final classifier as shown as white regions in the activation group differences. As with the mean activation maps, subsequent layers lose anatomic interpretability due to down sampling. This is further supported by the Class Activation Maps shown in Fig. 13. Notable regional differences between normal and dysplastic variants are shown with red arrows. The normal structural cohort shows a predilection for more defined posterior cerebellar hemispheres, while stronger discriminator power in the dysplastic cohort comes from less developed anterior cerebellar structures nearing the brainstem.

## Discussion

We have introduced a computational framework for the application of 3D Convolutional Neural Networks on a dataset with a small sample size, showing excellent performance using cross-validation for assessment of subcortical neonatal brain dysmaturation. Despite gain in popularity, neural networks can still be prohibitively difficult to implement in a large subset of classification tasks. While computationally accessible, large datasets and deep architectures still require powerful hardware and long computation times to learn predictive models or classifiers. More importantly, the accuracy of the classifier is highly dependent on the variance encoded within the training dataset. Intensity-based segmentation tasks have the advantage of dense training datasets, with relatively low complexity modeled by the feature sets. In contrast, structural morphology-based classification tasks, as presented in this study, suffer from sparse data and the need for features of higher complexity to model the structural intricacies of key brain structures, including the cerebral cortex and subcortical structures such as the hippocampus or cerebellum. As our logistic regression results show, the available simple volumetric and conventional metrics were insufficient for detecting these complex morphological patterns using a linear model. This does not prove that linear models are inherently insufficient, as a higher dimensional vector

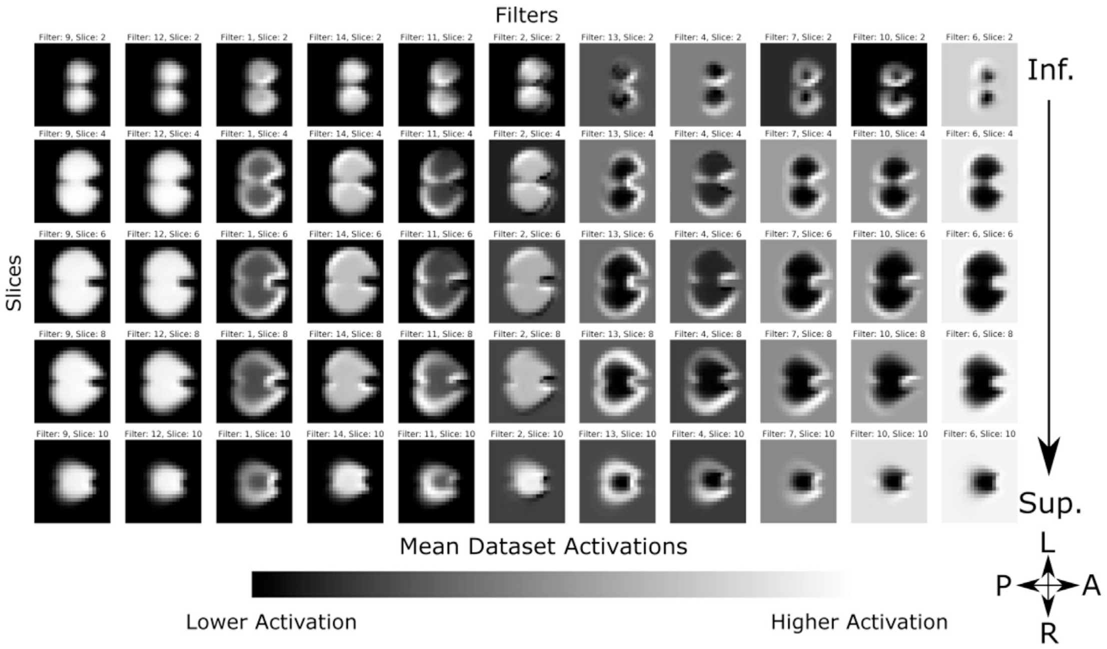

**Fig. 10.** Second convolutional layer mean activation. Filters are selectively sorted for visual clarity and intensities are scaled to show the contrast in activation range in each filter. Compared to the first layer, activations show more dramatic delineations of the superior surface of the cerebellum, co-localizing to the posterior subdivision of the bilateral cerebellar hemispheres and midline vermis.

using feature construction, or additional clinical or structural information may prove to be an adequate classifier using simpler linear models. However, more sophisticated non-linear approaches can also be explored using the structural information directly extracted from volumetric imaging.

For clinical populations, training sets are likely to be small, limiting the ability to apply 3D-CNN to more rare disorders. This is particularly difficult when implementing CNNs in medical imaging, as there can often be an insufficiently low incidence of abnormal cases in the training dataset, especially in neonatal and pediatric populations. These limitations can significantly slow down training, or impede it altogether, as the algorithm sees a disproportionate amount of normal cases through each iteration. Here we introduce a novel algorithm for neonatal brain dysmaturation that overcomes these challenges by first extracting the substructures of interest and registering them into a common space. This greatly reduces the computational search space and increases the observed effect size, leading to a relatively trivial application of deep neural networks as the final step. Most of the application of machine learning methods in neonatal neuroimaging to date have either been applied to actual segmentation of neonatal brain structures (Srhoj-Egekher et al., 2012; Song et al., 2007; Jaware et al., 2016) or functional resting BOLD imaging (Ball et al., 2016).

A secondary challenge encountered in deep learning is the inherently obfuscated nature of the features learned by the classifier. Classically, deep learning is considered a “black box”, where the set of features used by the algorithm are not known or interpretable by the user (Erhan et al., 2009). However, in medical imaging it is often desirable to generate mechanistic models of the phenotypical observations, rather than simply using a learned pattern to classify a disease. This necessitates the interpretability of the features learned within each hidden layer. Additionally, the coherence of the generated activation maps with known biological significance are added support for the specificity of structural morphology and subsequent clinical interpretation.

Although we have implemented a supervised learning method of classification, we can interpret the activation maps generated by the algorithm as data-driven learned features that provide insight into the underlying morphology of cerebellar dysplasia. The mean activation maps are effectively a lower dimensionality representation of the input data, with strength of activation in each group indicative of which regions in the image contribute most to the final classification task. Therefore, as we move further down each convolutional layer, the mean activations delineate further abstractions of the input images corresponding to the anatomical sub-regions of the cerebellum that are used for the final dysplasia classification. This technique is complementary to existing morphology based inference and classification methods, such as simple logistic regression, or deformation based statistical models (Cootes et al., 2018), including tensor based morphometry (Wang et al., 2011; Hua et al., 2008). Deformation based methods look at voxel-wise or point-mesh differences between groups based on deformation onto a common space, with statistical analysis performed on the parametrized deformation mapping variances across the set of images. These models have shown to be very powerful in the analysis of morphological variations in clinical populations, and generally require lower computational and dataset requirements. Such models may result in comparable classification accuracies, and may even be a more efficient model in comparison. Future work aimed at delineating the possible synergy between our proposed method and surface deformation methods is warranted. However, these methods are dependent on both the mesh and deformation algorithms used to compute their parameters, and introduce further model generation complexity. Therefore, for this present study, we chose to naively input the full structure into our network to prevent any bias from being introduced in a normalization or mesh-generating step. As part of our pre-processing, we only linearly register the input into a common space, but retain the native structure input. It was our expectation that as the algorithm learned the necessary features for classification, the weights associated with the interior structure would be reduced to zero, as can be seen in the activation maps. While this may be computationally more expensive during the learning phase, the end result is effectively comparable, and helps validate the anatomical relevance of the learned parameters. As such, not only can this technique be used for rigorous clinical translational mechanistic research studies, but can also be applied to clinical recognition of dysplastic abnormalities in individual neonatal patients.

Most deep learning and CNN applied to neuroimaging data up to date have not revealed insight into the neuropathologic underpinnings of the disease process studies (Vieira et al., 2017). In contrast, our algorithm

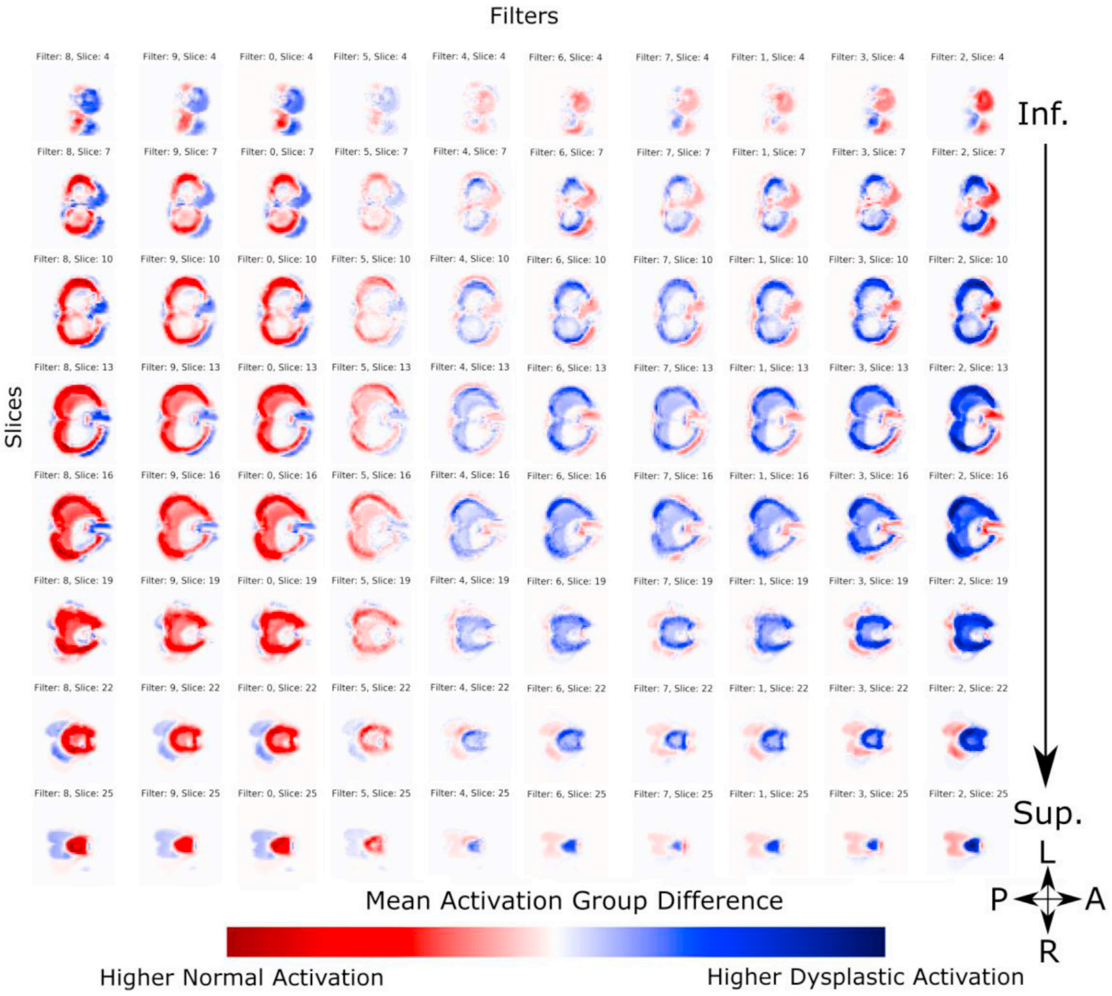

**Fig. 11.** First convolutional layer activation difference between control and dysplastic cerebelli. Blue regions are areas in which we see higher activation in the dysplastic cerebelli, and red is increased in the normal substructures. White regions showed low activations equally in both cohorts, and are thus uninformative to the classification task. Filters are selectively sorted for visual clarity to highlight the differences in filters that show higher activations in normal cerebelli contrasted with dysplastic inputs. Each filter shows a predilection for primarily classifying either normal or dysplastic cerebelli, however, some overlap is observed.

revealed important regional vulnerability of cerebellar dysplasia in CHD infants that has both clinical and biological relevance. Concordant with the criteria that the clinical pediatric neuroradiologists used to classify the cerebellar dysplasia based on outer structural morphology, deep cerebellar regions (including the flocculonodular lobe) were not informative to the final classifier, shown as white regions in the activation group differences and cold regions in the class activation maps, compared to the superior surface of the cerebellum which anatomically corresponded to the posterior cerebellar cortical hemispheres and midline vermis. This supports the potential use of our automated technique in identifying cerebellar dysplasia in infants with CHD in the clinical setting and suggests that shape-related morphological disturbances in patients with CHD are present. To date, most morphological studies in CHD have focused on volumetric measurements, showing decreased brain volumes both globally and regionally present at birth and in the first year of life (von Rhein et al., 2015; Ortinau et al., 2012), with volumetric abnormalities in specific brain structures being predictive of neurobehavior (Owen et al., 2014). Our regional cerebellar dysplasia is concordant with our recent work in fetal MRI which observed shape abnormalities in patients diagnosed with CHD (Wong et al., 2017), specifically in cerebellar vermis.

Our technique revealed important regional vulnerability of the cerebellum in CHD infants, with the superior surface (which includes the posterior subdivision and the midline vermis) providing greater discriminatory differences between dysplastic and normal appearing tissue. This finding is clinically relevant as clinical pediatric neuroradiology experts typically use the midline vermis and posterior aspect of the cerebellum to help discriminate between normal and dysplastic tissue, and could provide the basis by which our technique worked in the setting of a relatively small clinical sample size. This finding is also concordant with other genetic-based neuroimaging studies that have documented cerebellar dysplasia localized to the posterior cerebellar hemispheres and the midline vermis in other non-CHD patients, including autistic subjects (Wegiel et al., 2010). In addition, the vermis and posterior cerebellar hemispheres have been shown to be phenotypically abnormal in certain genetic animal models of cerebellar dysplasia including hypomorphic Foxc1 mutant mice that have both granule and Purkinje cell abnormalities (Haldipur et al., 2017). Interestingly, we have recently described cerebellar dysplasia in CHD infants correlated with ciliary motion assessed by nasal scrapes for ciliary motion video microscopy (Panigrahy et al., 2016). Brain tissue from human fetuses with Joubert syndrome, a well-known ciliopathy associated with cerebellar dysplasia, showed impaired Shh-dependent cerebellar development, particularly in the vermis and posterior subdivisions, also concordant with our findings (Aguilar et al., 2012). Similarly, regional cerebellar dysplasia, also involving the vermis and posterior subdivision is frequently seen in our mouse CHD mutants harboring cilia related mutations (Liu et al., 2017; Li et al., 2015). Together these findings suggest cilia defects may play an

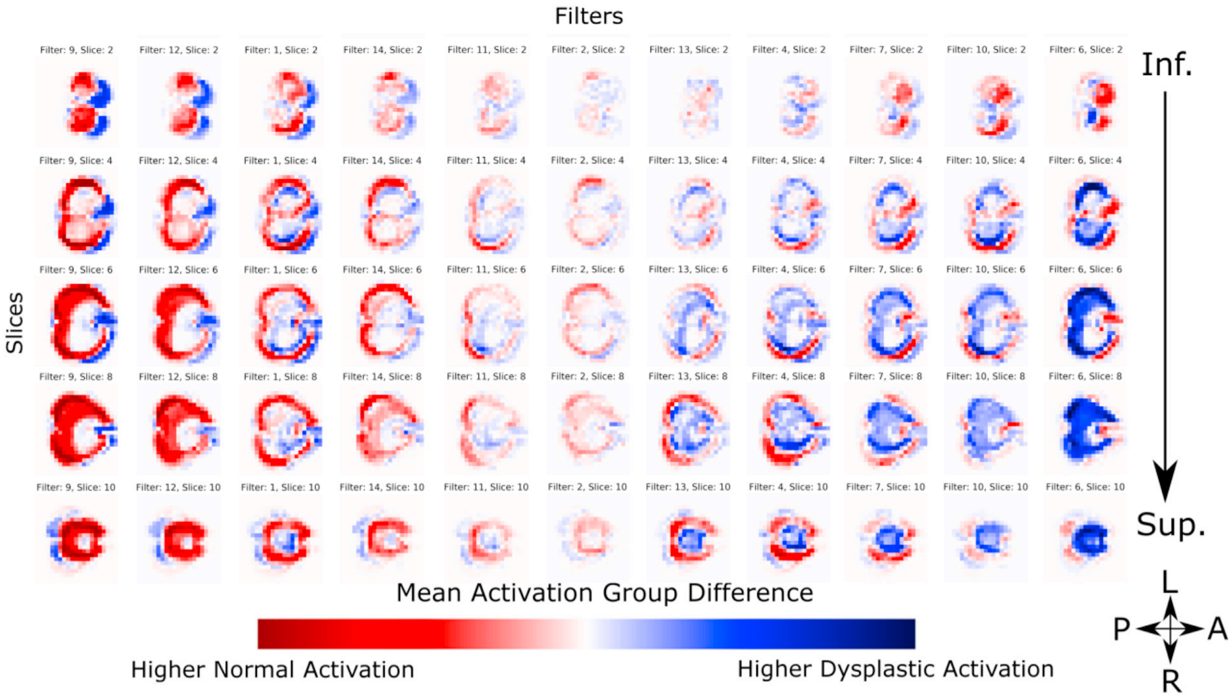

**Fig. 12.** Second convolutional layer activation difference between control and dysplastic cerebelli. Blue regions are areas in which we see higher activation in the dysplastic cerebelli, and red is increased in the normal substructures. White regions showed low activations equally in both cohorts, and are thus uninformative to the classification task. Filters are selectively sorted for visual clarity to highlight the differences in filters that show higher activations in normal cerebelli contrasted with dysplastic inputs. Compared to the first layer, we see more local delineation of regional cerebellar subdivisions. The superior surface of the cerebellum shows significant discriminatory power, seen as strongly differentiated regions of activation in the outer perimeter of the activation maps, which includes the posterior subdivision of the cerebellar lobes and the midline vermis.

important role in the pathogenesis of cerebellar dysplasia observed in CHD patients. In future work, we will aim to combine our deep learning techniques with our recently described novel computer vision approach for assessing the elemental ciliary motion parameters with the computation of optical flow in the digital videos (Quinn et al., 2015), to provide a more sensitive and robust phenotypic characterization of cerebellar dysplasia secondary to ciliary dysfunction.

The regional vulnerability of the CHD cerebellar dysplasia detects with our technique is likely to have downstream neurological and neurobehavioral consequences, which have been documented in CHD patients (Bellinger et al., 2011; Bellinger and Newburger, 2010). The primary function of the cerebellum has classically been associated with motor learning, muscle tone, coordination and language. However, more recent development indicates cerebellar modulation of higher cognitive functions, including working memory, processing speed, and executive functioning (D'Angelo and Casali, 2012). Eloquent cerebellar structure-function studies have demonstrated a topographic organization for motor regulation vs. cognitive and affective processing (Limperopoulos et al., 2010, 2014; Brossard-Racine et al., 2015; Allin et al., 2005); for example, Bolduc et al. showed decreased volume in the lateral cerebellar hemisphere was associated with impaired cognitive, expressive language, and motor control. While a reduction in vermal volume was related to impaired global development as well as behavior problems and a higher positive autism spectrum screening test (Bolduc et al., 2012). Furthermore, aberrant cerebellar development has been proposed to influence social and affect regulation, a term labeled cerebellar cognitive affective syndrome (CCAS) (Schmahmann and Sherman, 1998). These deficits have long term consequences, and early intervention is imperative (Calderon and Bellinger, 2015). Developing more sensitive methods for early detection of cerebellar impairment is an important step in the treatment and supportive neurodevelopmental care of infants with congenital heart disease. It is becoming more evident that localized injury to cerebellar subdivisions have varying effects in downstream neurodevelopment and behavior (Tiemeier et al., 2010). Current guidelines rely on qualitative observations that lack the sensitivity for more subtle morphological malformations that may have clinical impact in long-term development. By more clearly defining key regions of the cerebellum, which are vulnerable to dysmaturation, we can more aptly create objective diagnostic guidelines. Future works will examine the role of CNN detection of regional vulnerability of cerebellar dysplasia and neurobehavioral outcomes in CHD patients.

There are many other known clinical risk factors thought to contribute to cerebellar dysmaturation, which have been studied more extensively in premature infants, who also demonstrate different forms of brain dysmaturation, including cerebellar hypoplasia. Exposure to glucocorticoids, commonly used to accelerate lung maturation in-utero and to treat hypotension, is known to inhibit sonic hedgehog pathways (SHH) critical to cerebellar development, was not extensive used in our CHD patients (Back and Miller, 2014). Additionally, direct exposure to blood products and hemosiderin from cerebellar hemorrhage can have a direct effect on cerebellar growth (von Rhein et al., 2015). Finally, cerebral injury such as intra-ventricular hemorrhage (IVH) and more severe white matter injury (WMI) can lead to downstream disruption in cerebellar development (Tam, 2013). We excluded cases of cerebellar hemorrhage, IVH and WMI from our analysis in the CHD infants, so these are not likely contributing to our results. Future studies will be performed to examine the relationship of important perioperative clinical risk factors and CNN features of cerebellar dysplasia in CHD infants.

This study has several limitations. First, our sample size is small. The incidence of cerebellar dysplasia in this population is low, as there is a clear selection bias against more injured infants who may not have been deemed healthy enough to enroll in the study or have additional injuries that would exclude them, including gross ventriculomegaly and intraventricular hemorrhage. We did correct for pre-operative and post-operative status, and we did have few cases that had imaging in both periods. Another current limitation is the inclusion of a manual

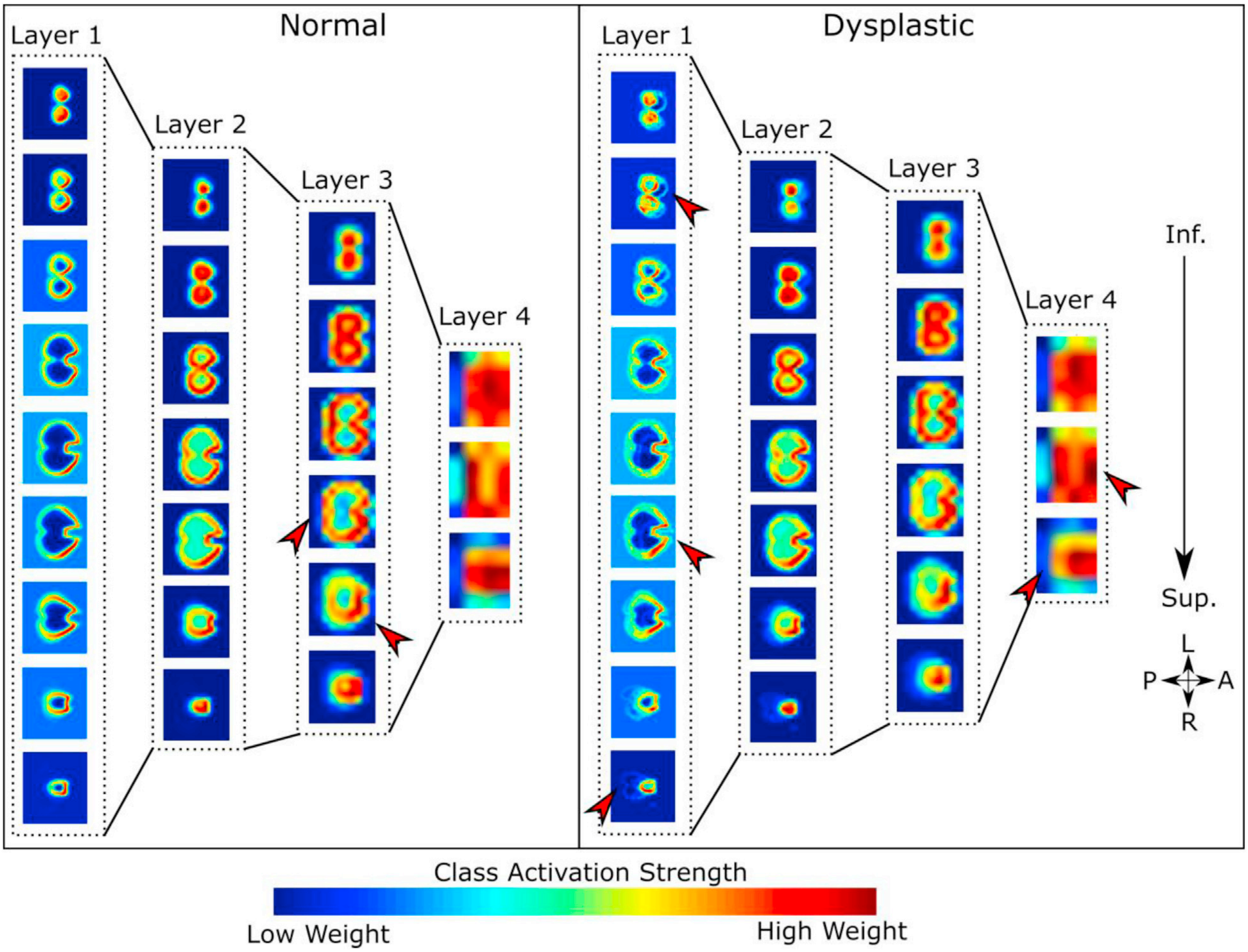

**Fig. 13.** Within-Class Average Activation Maps for normal and dysplastic cohorts. Within-Class Average Activation Maps (wCAMs) are constructed at each convolutional layer by calculating a weighted sum of the activations learned for each feature map at that layer. This generates a weighted spatial map across all feature maps that most contribute to the classifier at each layer. Since we have fully connected layers prior to the final classifier, we lose the direct connection of each individual feature map's contribution to the final classifier. However, it allows us to extract salient anatomical information at each convolutional layer. Notable regional differences between normal and dysplastic variants are shown with red arrows.

correction step to ensure accurate segmentation of the cerebellum prior to training the algorithm. This step may introduce user bias into the input, despite the users being blind to clinical diagnosis. Additionally, we validate our results using cross-validation. While cross-validation, along with the inclusion of drop-out layers during training, help prevent overtraining of the algorithm, it does not completely eliminate biases implicit to the dataset, such as population-specific variance and data acquisition parameters. Unfortunately, the use of a hold-out set given the available dataset significantly impacts the availability of abnormal structures used for training, hindering the learning process. As such, our results may still be overfitted. However, our framework's ability to discriminate subtle dysplastic cerebelli is encouraging, and the inferential results into the biological underpinnings of cerebellar dysmaturation are significant. In the future, we aim to test our method on an independently acquired dataset for external validation. We chose to train a 3D CNN, rather than use transfer learning of existing high-performing models such as AlexNet, to preserve the 3D structure of the input substructures. While using an existing feature rich model trained on a much larger dataset could result in more powerful discriminatory performance, it would require pre-processing of the input data that would diminish the interpretability of the structural morphology associated with the learned classifiers. Therefore, we believe that generating interpretable features is a worthwhile tradeoff. Our future work is geared towards fully automating the structure extraction step of the pipeline to ensure a completely user-independent throughput and broadening the scope to target additional brain structures and populations.

## Conclusion

We have introduced a computational framework for the extraction and specific classification of brain dysmaturation of subcortical structures in neonatal MRI, using a 3D convolutional neural network. We achieved a mean classification accuracy of 98.5% using 10-fold cross-validation. Furthermore, the hidden layer activations and class activation maps provide insight into the regional morphological characteristics of these dysplastic substructures in this at-risk population, which has both clinical diagnostic and biological implications. Our finding of the posterior lobe and vermis having the greatest discriminatory power for distinguishing dysplastic tissue from normal tissue in CHD infants, which is likely driven by genetic mechanisms. These findings also likely have important prognosticating implications with regard to neuromotor and neurobehavioral impairments in CHD patients. The code developed for this project is open source and published under the BSD License.

## Funding

This work was supported by the Department of Defense (W81XWH-16-1-0613), the National Heart, Lung, and Blood Institute (R01 HL128818-03), the National Institute of Neurological Disorders and Stroke (K23-063371), the U.S. National Library of Medicine

(5T15LM007059-27), the Pennsylvania Department of Health, the Mario Lemieux Foundation, and the Twenty Five Club Fund of Magee Women's Hospital.

## Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.05.049.

## References

Aguilar, A., Meunier, A., Strehl, L., et al., 2012. Analysis of human samples reveals impaired SHH-dependent cerebellar development in Joubert syndrome/Meckel syndrome. Proc. Natl. Acad. Sci. Unit. States Am. 109 (42), 16951–16956. https://doi.org/10.1073/pnas.1201408109.

Allin, M.P.G., Salaria, S., Nosarti, C., Wyatt, J., Rifkin, L., Murray, R.M., 2005. Vermis and lateral lobes of the cerebellum in adolescents born very preterm. Neuroreport 16 (16), 1821–1824. https://doi.org/10.1097/01.wnr.0000185014.36939.84.

Avants, B., Tustison, N., Song, G., 2009. Advanced normalization tools (ANTS). Insight J 1–35. ftp://ftp3.ie.freebsd.org/pub/sourceforge/a/project/ad/advants/Documentation/ants.pdf.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54 (3), 2033–2044. https://doi.org/10.1016/j.neuroimage.2010.09.025.

Back, S.A., Miller, S.P., 2014. Brain injury in premature neonates: a primary cerebral dysmaturation disorder? Ann. Neurol. 75 (4), 469–486. https://doi.org/10.1002/ana.24132.

Ball, G., Aljabar, P., Arichi, T., et al., 2016. Machine-learning to characterise neonatal functional connectivity in the preterm brain. Neuroimage 124 (Pt A), 267–275. https://doi.org/10.1016/j.neuroimage.2015.08.055.

Barkovich, A.J., Raybaud, C., 2012. Pediatric Neuroimaging. LWW.

Bellinger, D.C., Newburger, J.W., 2010. Neuropsychological, psychosocial, and quality-of-life outcomes in children and adolescents with congenital heart disease. Prog. Pediatr. Cardiol. 29 (2), 87–92. https://doi.org/10.1016/j.ppedcard.2010.06.007.

Bellinger, D.C., Wypij, D., Rivkin, M.J., et al., 2011. Adolescents with d-transposition of the great arteries corrected with the arterial switch procedure: neuropsychological assessment and structural brain imaging. Circulation 124 (12), 1361–1369. https://doi.org/10.1161/CIRCULATIONAHA.111.026963.

Bolduc, M.-E., du Plessis, A.J., Sullivan, N., et al., 2012. Regional cerebellar volumes predict functional outcome in children with cerebellar malformations. Cerebellum 11 (2), 531–542. https://doi.org/10.1007/s12311-011-0312-z.

Bosemani, T., Poretti, A., 2016. Cerebellar disruptions and neurodevelopmental disabilities. Semin. Fetal Neonatal Med. 1–10. https://doi.org/10.1016/j.siny.2016.04.014.

Brosch, T., Tang, L.Y.W., Yoo, Y., Li, D.K.B., Traboulsee, A., Tam, R., 2016. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE Trans. Med. Imag. 35 (5), 1229–1239. https://doi.org/10.1109/TMI.2016.2528821.

Brossard-Racine, M., du Plessis, A.J., Limperopoulos, C., 2015. Developmental cerebellar cognitive affective syndrome in ex-preterm survivors following cerebellar injury. Cerebellum 14 (2), 151–164. https://doi.org/10.1007/s12311-014-0597-9.

Calderon, J., Bellinger, D.C., 2015. Executive function deficits in congenital heart disease: why is intervention important? Cardiol. Young (January), 1–9. https://doi.org/10.1017/S1047951115001134.

Clevert, D.-A., Unterthiner, T., Hochreiter, S., November 2015. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). https://doi.org/10.3233/978-1-61499-672-9-1760.

Cootes T.F., Twining C.J., Taylor C.J. Diffeomorphic Statistical Shape Models. http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/Papers/cootes_bmvc04.pdf. Accessed 8 January 2018.

Doherty, D., Millen, K.J., Barkovich, A.J., 2013. Midbrain-hindbrain malformations: advances in clinical diagnosis, imaging, and genetics. Lancet Neurol. 12 (4), 381–393. https://doi.org/10.1016/S1474-4422(13)70024-3.

D'Angelo, E., Casali, S., 2012. Seeking a unified framework for cerebellar function and dysfunction: from circuit operations to cognition. Front. Neural Circ. 6 (January), 116. https://doi.org/10.3389/fncir.2012.00116.

Erhan, D., Bengio, Y., Courville, A., Vincent, P., 2009. Visualizing higher-layer features of a deep network. Bernoulli (1341), 1–13. https://www.researchgate.net/profile/Aaron_Courville/publication/265022827_Visualizing_Higher-Layer_Features_of_a_Deep_Network/links/53ff82b00cf24c81027da530.pdf. (Accessed 27 June 2017).

Gorgolewski, K., Burns, C.D., Madison, C., et al., 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front. Neuroinf. 5 https://doi.org/10.3389/fninf.2011.00013.

Gousias, I.S., Edwards, A.D., Rutherford, M.A., et al., 2012. Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. Neuroimage 62 (3), 1499–1509. https://doi.org/10.1016/j.neuroimage.2012.05.083.

Gousias, I.S., Hammers, A., Counsell, S.J., et al., 2013. Magnetic resonance imaging of the newborn brain: automatic segmentation of brain images into 50 anatomical regions. In: Rodriguez-Fornells, A. (Ed.), PLoS One, vol. 8, e59990. https://doi.org/10.1371/journal.pone.0059990, 4.

Gupta A., Ayhan M.S., Maida A.S. Natural Image Bases to Represent Neuroimaging Data. http://proceedings.mlr.press/v28/gupta13b.pdf. Accessed 4 October 2017.

Haldipur, P., Dang, D., Aldinger, K.A., et al., 2017. Phenotypic outcomes in Mouse and Human Foxc1 dependent Dandy-Walker cerebellar malformation suggest shared mechanisms. Elife 6, 1–15. https://doi.org/10.7554/eLife.20898.

Hosseini-Asl, E., Gimel'farb, G., El-Baz, A., Gimel 'farb, G., El-Baz, A., 2016. Alzheimer's Disease Diagnostics by a Deeply Supervised Adaptable 3D Convolutional Network, 502. https://doi.org/10.1109/ICIP.2016.7532332.

Hua, X., Leow, A.D., Parikshak, N., et al., 2008. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. Neuroimage 43 (3), 458–469. https://doi.org/10.1016/j.neuroimage.2008.07.013.

Jaware, T.H., Khanchandani, K.B., Zurani, A., 2016. Multi-kernel support vector machine and Levenberg-Marquardt classification approach for neonatal brain MR images. In: 2016 IEEE 1st Int Conf Power Electron Intell Control Energy Syst, vol. 8, pp. 1–4. https://doi.org/10.1109/ICPEICES.2016.7853639, 0.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. FSL. Neuroimage 62 (2), 782–790.

Kleesiek, J., Urban, G., Hubert, A., et al., 2016. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. Neuroimage 129, 460–469. https://doi.org/10.1016/j.neuroimage.2016.01.024.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2323. https://doi.org/10.1109/5.726791.

Li, Y., Klena, N.T., Gabriel, G.C., et al., 2015. Global genetic analysis in mice unveils central role for cilia in congenital heart disease. Nature 521 (7553), 520–524. https://doi.org/10.1038/nature14269.

Limperopoulos, C., Chilingaryan, G., Guizard, N., et al., 2010. Cerebellar injury in the premature infant is associated with impaired growth of specific cerebral regions. Pediatr. Res. 68 (2), 145–150. https://doi.org/10.1203/00006450-201011001-00282.

Limperopoulos, C., Chilingaryan, G., Sullivan, N., Guizard, N., Robertson, R.L., Du Plessis, A.J.A.J.A.J., 2014. Injury to the premature cerebellum: outcome is related to remote cortical development. Cerebr. Cortex 24 (3), 728–736. https://doi.org/10.1093/cercor/bhs354.

Liu, X., Yagi, H., Saeed, S., et al., 2017. The complex genetics of hypoplastic left heart syndrome. Nat. Genet. 49 (7), 1152–1159. https://doi.org/10.1038/ng.3870.

Mueller, S.G., Weiner, M.W., Thal, L.J., et al., 2005. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). Alzheimer's Dementia 1 (1), 55–66. https://doi.org/10.1016/j.jalz.2005.06.003.

Nielsen MA. Neural Networks and Deep Learning, 2015. Determination Press. http://neuralnetworksanddeeplearning.com/.

Oegema, R., Cushion, T.D., Phelps, I.G., et al., 2015. Recognizable cerebellar dysplasia associated with mutations in multiple tubulin genes. Hum. Mol. Genet. 24 (18), 5313–5325. https://doi.org/10.1093/hmg/ddv250.

Ortinau, C., Inder, T., Lambeth, J., Wallendorf, M., Finucane, K., Beca, J., 2012. Congenital heart disease affects cerebral size but not brain growth. Pediatr. Cardiol. 33 (7), 1138–1146. https://doi.org/10.1007/s00246-012-0269-9.

Owen, M., Shevell, M., Donofrio, M., et al., 2014. Brain volume and neurobehavior in newborns with complex congenital heart defects. J. Pediatr. 164 (5) https://doi.org/10.1016/j.jpeds.2013.11.033, 1121–1127.e1.

Panigrahy, A., Lee, V., Ceschin, R., et al., 2016. Brain dysplasia associated with ciliary dysfunction in infants with congenital heart disease. J. Pediatr. 178 https://doi.org/10.1016/j.jpeds.2016.07.041, 141–148.e1.

Payan, A., Montana, G., 2015. Predicting Alzheimer's Disease: a Neuroimaging Study with 3D Convolutional Neural Networks, pp. 1–9. https://doi.org/10.1613/jair.301.

Poretti, A., Boltshauser, E., 2016. Huisman TAGM. Pre- and postnatal neuroimaging of congenital cerebellar abnormalities. Cerebellum 15 (1), 5–9. https://doi.org/10.1007/s12311-015-0699-z.

Quinn, S.P., Zahid, M.J., Durkin, J.R., Francis, R.J., Lo, C.W., Chennubhotla, S.C., 2015. Automated identification of abnormal respiratory ciliary motion in nasal biopsies. Sci. Transl. Med. 7 (299), 299ra124. https://doi.org/10.1126/scitranslmed.aaa1233.

Rajpurkar, P., Irvin, J., Zhu, K., et al., 2017. CheXNet: Radiologist-level Pneumonia Detection on Chest X-rays with Deep Learning, pp. 3–9 doi:1711.05225.

Schmahmann, J.D., Sherman, J.C., 1998. The cerebellar cognitive affective syndrome. Brain 121 (4), 561–579. https://doi.org/10.1093/brain/121.4.561.

Song, Z., Awate, S.P., Licht, D., Gee, J.C., 2007. Clinical neonatal brain MRI segmentation and intensity-based markov priors. Computer 1–8. https://doi.org/10.1007/978-3-540-75757-3_107.

Srhoj-Egekher, V., Benders, M., Kersbergen, K.J., Viergever, M.A., Isgum, I., 2012. Automatic segmentation of neonatal brain MRI using atlas based segmentation and machine learning approach. MICCAI Gd Chall Neonatal Brain Segmentation.

Stoodley, C.J., Limperopoulos, C., 2016. Structure–function relationships in the developing cerebellum: evidence from early-life cerebellar injury and neurodevelopmental disorders. Semin. Fetal Neonatal Med. 21 (5), 1–9. https://doi.org/10.1016/j.siny.2016.04.010.

Szegedy, C., Zaremba, W., Sutskever, I., et al., December 2013. Intriguing Properties of Neural Networks. https://doi.org/10.1021/ct2009208.

Tam, E.W.Y., 2013. Potential mechanisms of cerebellar hypoplasia in prematurity. Neuroradiology 55 (Suppl. 2), 41–46. https://doi.org/10.1007/s00234-013-1230-1.

Theano Development Team, 2016. Theano: a Python Framework for Fast Computation of Mathematical Expressions, vol 19. arXiv e-prints. http://arxiv.org/abs/1605.02688.

Tiemeier, H., Lenroot, R.K., Greenstein, D.K., Tran, L., Pierson, R., Giedd, J.N., 2010. Cerebellum development during childhood and adolescence: a longitudinal morphometric MRI study. Neuroimage 49 (1), 63–70. https://doi.org/10.1016/j.neuroimage.2009.08.016.

Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage 145, 166–179. https://doi.org/10.1016/J.NEUROIMAGE.2016.10.038.

Vieira, S., Pinaya, W.H.L., Mechelli, A., 2017. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. Neurosci. Biobehav. Rev. 74, 58–75. https://doi.org/10.1016/j.neubiorev.2017.01.002.

Volpe, J.J., 2014. Encephalopathy of Congenital Heart Disease- Destructive and Developmental Effects Intertwined. https://doi.org/10.1016/j.jpeds.2014.01.002.

von Rhein, M., Buchmann, A., Hagmann, C., et al., 2015. Severe congenital heart defects are associated with global reduction of neonatal brain volumes. J. Pediatr. 167 (6) https://doi.org/10.1016/j.jpeds.2015.07.006, 1259–1263.e1.

Vovk, U., Pernuš, F., Likar, B., 2007. A review of methods for correction of intensity inhomogeneity in MRI. IEEE Trans. Med. Imag. 26 (3), 405–421. https://doi.org/10.1109/TMI.2006.891486.

Wagner R, Thom M, Schweiger R, Palm G, Rothermel A. Learning Convolutional Neural Networks From Few Samples. http://geza.kzoo.edu/~erdi/IJCNN2013/HTMLFiles/PDFs/P274-1108.pdf. Accessed 8 January 2018.

Wang, Y., Panigrahy, A., Shi, J., et al., September 2011. Surface multivariate tensor-based morphometry on premature neonates: a pilot study. In: 2nd MICCAI 2013 Work Clin Image-based Proced Transl Res Med Imaging.

Wegiel, J.J., Kuchna, I., Nowicki, K., et al., 2010. The neuropathology of autism: defects of neurogenesis and neuronal migration, and dysplastic changes. Acta Neuropathol. 119 (6), 755–770. https://doi.org/10.1007/s00401-010-0655-4.

Wong, A., Chavez, T., O 'neil, S., et al., 2017. Synchronous aberrant cerebellar and opercular development in fetuses and neonates with congenital heart disease: correlation with early communicative neurodevelopmental outcomes, initial experience. Am. J. Perinatol. Rep. 7 (1), 17–27. https://doi.org/10.1055/s-0036-1597934.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. Comput Vision–ECCV 8689, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53.

Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R., 2010. Deconvolutional networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2528–2535. https://doi.org/10.1109/CVPR.2010.5539957.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imag. 20 (1), 45–57. https://doi.org/10.1109/42.906424.

Zhao, H., Lu, Z., Poupart, P., 2015. Self-adaptive hierarchical sentence model. In: IJCAI Int Jt Conf Artif Intell, pp. 4069–4076. https://doi.org/10.1162/153244303322533223, 2015-Janua.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., December 2015. Learning Deep Features for Discriminative Localization. https://doi.org/10.1109/CVPR.2016.319.

Zou, K.H., Warfield, S.K., Bharatha, A., et al., 2004. Statistical validation of image segmentation quality based on a spatial overlap index. Acad. Radiol. 11 (2), 178–189. https://doi.org/10.1016/S1076-6332(03)00671-8.